Book Review


The problem of observed-score equating has drawn the attention of test theorists for nearly a century. It may seem difficult to write a new book on such an old topic but the authors of this book managed to do so. Thus far, test equating mainly consisted of a collection of separate techniques for distribution estimation, smoothing of sample estimates, score transformation, and the like. The authors have succeeded in integrating these techniques in a new framework, which they call kernel equating. Kernel equating has been in use at Educational Testing Service (ETS) for some time; this book paves the way for those outside of ETS with an interest in this framework.

The book is organized as an introductory chapter and two five-chapter sections, one on theory and the other on applications. It also has appendices that introduce the δ-method, the theory of fitting a log-linear model to a sample distribution, and give a review of the matrix notation and operations used in this book.

The introductory chapter defines the equating problem, introduces the notation, discusses the two main equating functions used in this book (linear and equipercentile functions), and prepares the statistical aspects of observed-score equating (choice of sampling design and definition of standard error of equating). The chapter also touches on the five requirements that are generally considered as the foundations of observed-score equating (common construct; equal reliability; symmetry; equity; and population invariance). The chapter does not describe how these requirements relate to the type of equating addressed throughout this book (equating of score distributions on two different tests for a target population of test takers).

Chapter 2 introduces the data collection designs commonly used for observed-score equating: the equivalent-groups (EG) design, single-group (SG) design, counterbalanced (CB) design, and the nonequivalent-groups-with-anchor-test (NAET) design. This chapter can be recommended for use as an introductory text in a class on test equating. The treatment of each of these designs is clear and precise. I especially liked the discussion of the compound nature of the counterbalanced and the two alternative treatments of the data collected in the NAET design (chain equating and post-stratification equating). A new concept introduced in this chapter is that of a design function. This function maps the population distributions for the two tests into the probability distributions of a target population for a given design. The function takes a simple form for the equivalent-groups and single-group designs but requires more complicated matrix notation for the other designs.

The first step of kernel equating is presmoothing of the raw observed-score data. The authors choose log-linear smoothing and their choice is explained and motivated in Chapter 3. Linear models for the log probabilities of (discrete) observed scores
allow us to fit observed-score distributions with any desired degree of precision, including distributions with aberrancies that are typical of some test scores. They also allow us to calculate the (large-sample) covariance matrix of the MLEs of their parameters. This chapter is the first introduction to this topic in a test-theory textbook and is also recommended for class use.

Chapter 4 is on the continuization of the (smoothed) distributions for the target population. Continuization is necessary because it is generally impossible to map one discrete observed-score distribution onto another preserving all percentile ranks. The equating framework presented in this book derives its name from the authors’ choice of kernel smoothing for this step. The smoothed versions of the two population distributions are used in the equating function. The use of a Gaussian kernel is a key step in the process. The choice of a larger bandwidth makes the score distributions approximate a normal form and results in a linear equating function. Except for this limiting case, we have a regular equipercentile function. The authors offer a penalty function to assist in the choice of bandwidth.

Chapter 5 treats the final step in the kernel equating method—the evaluation of the equating transformation. The chapter contains a general expression for the (large-sample) standard error of equating for the kernel method that is derived using the δ-method. The expression is a nice composition of three quantities, each of which is related to a different step in the equating process: (i) the (large-sample) covariance matrix of the log-linear smoothing of the raw data, (ii) a Jacobean due to the design function used to transform the smoothed probabilities into those for the target population, and (iii) a Jacobean to account for the equating function. The second component is dependent on the data collection design and the authors detail its computation for the various designs treated in the book. Another attractive result derived in this chapter is the (large-sample) standard error of the difference between kernel equating functions. This standard error can be used, for instance, to evaluate the differences between the results for the two alternative methods for equating for the NAET design (chain equating and post-stratification equating).

In Chapter 6, the kernel method of equating is compared both with equipercentile equating that exploits linear score interpolation to make the scores continuous (“percentile rank method”) and linear equating. The comparison is facilitated by the fact that the former can be viewed as kernel smoothing with a uniform kernel.

Chapters 7–11 give examples of kernel equating for each of the data collection designs treated in this book. These chapters are most attractive in that they allow us to look over the authors’ shoulders when they actually apply kernel equating to real test data. It is particularly instructive to watch when the authors try alternatives, for instance, for log-linear presmoothing. The reader gains from understanding why the authors make their choices. Also, the ample use of figures with fitted and observed results and bands for the standard errors give the reader a valuable sense of the shape of the equating functions and the sizes of the errors in a typical example. The only step that is not explained in the book is how the authors used their penalty function for determining the bandwidth for the kernel smoother in these applications. (But one of the authors has told me that they use a method known as Brent’s method and that software for it is available from dlembacc@ets.org.)
Another interesting portion of the book is the empirical comparison between chain equating and post-stratification equating for the NAET design in Chapters 10 and 11. Post-stratification resulted in significantly larger equated scores. Although the differences were small, they are large enough to give rise to a discussion in high-stakes testing. Since the two methods rest on choices that have no empirical justification, the authors suggest analyzing their robustness to various test and population quantities as a criterion of choice.

*The Kernel Method of Test Equating* takes the tradition of observed-score equating in testing to a higher level of sophistication. It has been written in a style that nicely balances lucidity and practical relevance. I heartily recommend this book to graduate students and practitioners with an interest in this tradition. As already indicated, except for some of the more technical material in Chapters 4 and 5, the book will be an excellent textbook for a graduate class on test equating. The comments that I have are not on the book but on the current tradition of observed-score equating to which the book belongs.

First, I am puzzled by the dominant role in the theory of test equating that is played by the continuization of an inherently discrete distribution of, say, a number-correct score on a test. The book has convinced me that the choice of type of continuization is not entirely arbitrary; it makes sense to preserve the most important moments of the discrete distribution. But, still, it seems somewhat embarrassing that the size of the bandwidth for the smoother implies the choice between a linear and a curved family of equating functions. It would have been more convincing if the choice depended on the third- and higher-order moments of the empirical distributions for the two tests.

Second, the tradition of observed-score equating seems to focus almost exclusively on the (large-sample) standard error of equating as a criterion of success. The possibility of bias in test equating is ignored. For example, it is well known that the choice of a bandwidth in kernel smoothing is liable to a bias-accuracy tradeoff. Since the bias is included in the component with the squared-error component (Eq. 4.27) of the penalty function used to choose the bandwidth in this book, the equating function, which is based on two kernel-smoothed distributions, is also expected to be biased. Why not use the impact of the actual size of this component on the final equating function as an additional criterion?

Finally, the only reason why the current tradition of equating treats observed scores as random variables is that they are considered to be for randomly selected examinees from some population (p. 6). The existence of measurement error as another—and in my view more convincing—source of randomness is ignored. The only appropriate way to allow for measurement error is by modeling the observed scores with a random component. Both classical and modern test theory are based on this principle. If measurement error is ignored, we are bound to pay a price in the form of bias in our results (here: equated scores). In a review of the first edition of the book on test equating by Kolen and Brennan (1995), I had to make the same comment and observed that it was “...time for test equating to get a firm psychometric footing.” (van der Linden, 1997). The observation still seems to hold.

Wim J. van der Linden
*University of Twente*
References


Author

WIM J. VAN DER LINDEN is a Professor of the Department of Research Methodology, Measurement, and Data Analysis, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands; w.j.vanderlinden@utwente.nl. His primary research interests include test theory, applied statistics, and research methods.